Note: natural logarithms are used throughout, unless stated otherwise.

1. **VC classes.** Prove the following statements.

   (a) Let $\mathcal{C}$ and $\mathcal{C}'$ be two classes of subsets of some feature space $X$. Suppose that $\mathcal{C} \subseteq \mathcal{C}'$, meaning that if $C \in \mathcal{C}$, then $C \in \mathcal{C}'$ as well. Prove that $V(\mathcal{C}) \leq V(\mathcal{C}')$.

   (b) Let $\mathcal{C}$ be a finite class of subsets of $X$. Prove that $V(\mathcal{C}) \leq \log_2 |\mathcal{C}|$.

   (c) Let $X$ be a finite feature space. For a given $k \leq |X|$, consider the class $\mathcal{F}_k$ of binary-valued functions $f : X \rightarrow \{0,1\}$, such that $|\{x \in X : f(x) = 1\}| = k$. Find $V(\mathcal{F}_k)$.

2. **An alternative to ERM.** Searching for an empirical risk minimizer in an infinite function class $\mathcal{F}$ may not always be feasible. Let’s consider the following alternative procedure that reduces to searching over a finite subclass of $\mathcal{F}$. We will be looking at a binary classification problem, so let $\mathcal{F}$ be a class of functions $f : X \rightarrow \{0,1\}$, where $X$ is some feature space. Given a training sample $\{(X_i, Y_i)\}_{i=1}^{n}$ from an unknown probability distribution $P$ on $X \times \{0,1\}$, we carry out the following two-step procedure:

   • Pick some $m < n$. Let $B_m$ be the set of all binary strings in $\{0,1\}^m$ of the form
     
     \[ b^m(f) = (b_1(f), \ldots, b_m(f)) = (f(X_1), \ldots, f(X_m)) \]  

     for some $f \in \mathcal{F}$. For each $b \in B_m$, pick one $f \in \mathcal{F}$ such that (1) holds. Let $\hat{\mathcal{F}}_m$ denote the (finite) set of all such $f$’s. Note that this is a random set, since it depends on the sample $(X_1, \ldots, X_m)$.

   • Compute

     \[ \hat{f}_n = \arg\min_{f \in \hat{\mathcal{F}}_m} \frac{1}{n-m} \sum_{i=m+1}^{n} 1_{\{f(X_i) \neq Y_i\}}. \]

     This will be our actual classifier.

What we have done is split the original training sample into two subsamples, used the first subsample to extract a finite subclass of $\mathcal{F}$, and then performed ERM over this subclass on the second subsample. We will now analyze the classification error of $\hat{f}_n$. 

(a) Let
\[ \hat{f}_m = \arginf_{f \in \hat{F}_m} L(f) \]
be the best classifier in \( \hat{F}_m \). Note that this is a random object, since it depends on the random set \( \hat{F}_m \). Prove that
\[ L(\hat{f}_n) - L(\hat{f}_m) \leq 8\sqrt{\frac{\log |\mathcal{S}_m(\mathcal{F})|}{n-m}} + \sqrt{\frac{2\log(2/\delta)}{n-m}} \]  \hspace{1cm} (2)
with probability at least \( 1 - \delta/2 \), where \( \mathcal{S}_m(\mathcal{F}) \) is the \( m \)th shatter coefficient of \( \mathcal{F} \).

\textbf{Hint:} Use the fact that \( \hat{f}_n \) is a solution of an ERM problem over the second subsample, add and subtract appropriate empirical quantities, and then apply the Finite Class Lemma.

(b) Observe that
\[ L(\hat{f}_m) - L^*(\mathcal{F}) \leq \sup_{f, f' \in \mathcal{F}: b_m(f) = b_m(f')} \left| L(f) - L(f') \right| \]
\[ \leq \sup_{f, f' \in \mathcal{F}: b_m(f) = b_m(f')} \mathbb{P} \left[ f(X) \neq f'(X) \right] \]
\[ \leq \sup_{A \in \mathcal{A}} \left| \mathbb{P}(X \in A) - \frac{1}{m} \sum_{i=1}^{m} 1_{\{X_i \in A\}} \right| , \]
where \( \mathcal{A} \) is the class of all sets of the form \( \{x \in X : f(x) \neq f'(x)\} \) for all pairs \( f, f' \in \mathcal{F} \). Use this to prove that
\[ L(\hat{f}_m) - L^*(\mathcal{F}) \leq C \sqrt{\frac{V(\mathcal{A})}{m} + \frac{2\log(2/\delta)}{m}} \]
with probability at least \( 1 - \delta/2 \), where \( C > 0 \) is an absolute constant. (It is not hard to show that \( V(\mathcal{A}) \leq 4V(\mathcal{F}) \) — you don’t have to do this.)

(c) Finally, use parts (a)–(b) to prove that
\[ L(\hat{f}_n) - L^*(\mathcal{F}) \leq 8\sqrt{\frac{\log |\mathcal{S}_m(\mathcal{F})|}{n-m}} + C \sqrt{\frac{V(\mathcal{A})}{m}} + \sqrt{\frac{2\log(2/\delta)}{n-m}} + \sqrt{\frac{2\log(2/\delta)}{m}} \]
with probability at least \( 1 - \delta \).

3. \textbf{Surrogate loss bound for a sigmoidal classifier.} Let the feature space \( X \) be a subset of \( \mathbb{R}^d \). Consider the class \( \mathcal{F}_R \) of functions \( f \) the form
\[ f_w(x) \triangleq \tanh \left( \langle w, x \rangle \right) , \]  \hspace{1cm} (3)
where \( w \) runs over all vectors in \( \mathbb{R}^d \) with \( \|w\| \leq R \). Each such \( f \) induces a classifier \( g_f(x) = \text{sgn} f(x) \). A classifier of this kind first computes a weighted sum of the features, then passes it through a smooth nonlinear function, and then computes the sign of the resulting value. The hyperbolic tangent is an example of a \textit{sigmoidal function} (where “sigmoidal” is a fancy term for “S-shaped” — look at the graph of \( u \rightarrow \tanh u \)). The transformation in (3) is a simple model of a nonlinear neuron.
Let \( \varphi \) be a surrogate loss function satisfying the assumptions of Theorem 3 in the lecture notes on Binary Classification, Part 1. Let \( \hat{f}_n \in \mathcal{F}_L \) be a function generated by an arbitrary learning algorithm on the basis of an i.i.d. sample \( \{(X_i, Y_i)\}_{i=1}^n \) from an unknown probability distribution \( P \) on \( X \times \{-1, +1\} \). Prove that

\[
L(\hat{f}_n) \leq A_{\varphi, n}(\hat{f}_n) + 8RM_{\varphi} \sqrt{\frac{\mathbb{E}[\|X\|^2]}{n}} + B \sqrt{\frac{\log(1/\delta)}{2n}}
\]

with probability at least \( 1 - \delta \), where \( M_{\varphi} \) and \( B \) are defined in Theorem 3. Recall that \( L(f) \) is our shorthand for the error probability \( \mathbb{P}[\text{sgn}f(X) \neq Y] \).